Requested Patent: FR2821951A1

Title:

TOOL FOR AUTOMATICALLY HIGHLIGHTING DISTANCING LEVEL IN TEXTUAL DOCUMENTS, COMPRISES MEANS TO DETERMINE NUMBER OF CLAUSES WHICH CONTAIN EITHER HYPOTHETICAL CONDITIONAL OR CONDITIONAL PRESENT TENSES ;

Abstracted Patent: FR2821951 ;

Publication Date: 2002-09-13 ;

Inventor(s): GAY LOUIS;; MASSIOT OLIVIER ;

Applicant(s): DATOPS SA (FR) ;

Application Number: FR20010003015 20010306 ;

Priority Number(s): FR20010003015 20010306 ;

IPC Classification: G06F19/00; G06F17/30 ;

Equivalents: ;

ABSTRACT:

The number of clauses in a document which contain conditional hypothetical or present conditional without condition tenses is automatically determined and used to produce a rate of distancing per sentence. In a database evolving in time with new documents averages and standard deviations of the rates of distancing at successive instants are detected when they exceed predetermined thresholds

(19) RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL
DE LA PROPRIÉTÉ INDUSTRIELLE

PARIS

(11) N° de publication : **2 821 951**
(à n'utiliser que pour les commandes de reproduction)

(21) N° d'enregistrement national : **01 03015**

(51) Int Cl[7] : G 06 F 19/00, G 06 F 17/30

(12) DEMANDE DE BREVET D'INVENTION A1

(22) Date de dépôt : 06.03.01.

(30) Priorité :

(43) Date de mise à la disposition du public de la demande : 13.09.02 Bulletin 02/37.

(56) Liste des documents cités dans le rapport de recherche préliminaire : *Ce dernier n'a pas été établi à la date de publication de la demande.*

(60) Références à d'autres documents nationaux apparentés :

(71) Demandeur(s) : *DATOPS SA* — FR.

(72) Inventeur(s) : GAY LOUIS et MASSIOT OLIVIER.

(73) Titulaire(s) :

(74) Mandataire(s) : REGIMBEAU.

(54) OUTIL POUR LA MISE EN EVIDENCE AUTOMATIQUE D'UN NIVEAU DE DISTANCIATION DANS UN ENSEMBLE DE DONNEES TEXTUELLES.

(57) Outil pour le traitement d'au moins une partie d'un ensemble de documents textuels stockés dans une base de données, caractérisé en ce qu'il comporte des moyens pour déterminer automatiquement dans au moins un document le nombre de propositions dans lesquelles le temps conditionnel hypothétique et le nombre de propositions dans lesquelles le temps conditionnel présent sans condition et employé, et pour calculer un taux de distanciation qui est fonction du ou des nombres ainsi déterminés.

FR 2 821 951 - A1

OUTIL POUR LA MISE EN EVIDENCE AUTOMATIQUE D'UN NIVEAU DE DISTANCIATION DANS UN ENSEMBLE DE DONNEES TEXTUELLES

La présente invention est relative à un outil pour le traitement d'un ensemble de données textuelles.

Des outils pour la mise en œuvre de traitements automatiques sur des données textuelles sont déjà connus.

Notamment, il a déjà été proposé par la demanderesse dans sa demande de brevet WO99/05614 un outil permettant un suivi dynamique dans le temps de l'information contenue dans les documents d'une base de données. Cet outil permet en particulier de mesurer à plusieurs instants successifs un certain nombre de paramètres caractérisant les données textuelles contenues dans les documents et de mettre en évidence une éventuelle évolution anormale du contenu informatif de ceux-ci dans le temps.

Il a également été proposé par la demanderesse dans sa demande de brevet FR 00/11068 un procédé de traitement apte à permettre une détection particulièrement fiable de distorsions de la structure de l'information contenue dans les documents d'une base de données qui évolue dans le temps. Ce procédé permet de mettre en évidence très rapidement de telles distorsions, alors même que celles-ci seraient difficilement détectables par une lecture humaine.

L'invention a quant à elle pour but de mettre en évidence le niveau de distanciation dans un document, ceci afin de permettre de distinguer les textes dans lesquels le locuteur s'éloigne d'une affirmation directe et met une distance entre son opinion personnelle et le contenu de son discours. Le locuteur introduit le doute sur la situation réelle du sujet évoqué. Cette distanciation s'effectue par l'usage de temps et de tournures conditionnelles.

Pour mettre en évidence de façon automatique un tel niveau de distanciation, l'invention propose un outil pour le traitement d'au moins une partie d'un ensemble de documents textuels stockés dans une base de données, caractérisé en ce qu'il comporte des moyens pour déterminer automatiquement dans au moins un document le nombre de propositions dans lesquelles le temps conditionnel hypothétique est employé et le nombre de propositions dans lesquelles le temps conditionnel présent sans condition et employé, et pour calculer un taux de distanciation qui est fonction du ou des nombres ainsi déterminés.

D'autres caractéristiques et avantages de l'invention ressortiront encore de l'exemple qui va maintenant être décrit. Selon cet exemple, on acquiert un nombre important de documents constitués de données textuelles, par exemple en mettant en œuvre une recherche au moyen d'un moteur de recherche sur Internet ou encore en utilisant des bases de données spécifiques.

Ces documents sont mémorisés dans une base de données, qui est par exemple mise à jour régulièrement, de sorte que son contenu évolue dans le temps.

L'outil met en œuvre sur les documents de cette base de données différents traitements, par exemple les traitements décrits dans les demandes de brevet de la demanderesse WO99/05614 et FR 00/11068.

Il met également en œuvre pour chaque document un traitement d'analyse syntaxique, qui s'inspire des publications de Gosselin (« Sémantique de la temporalité en français. », 1996 ; et « La valeur de l'imparfait et du conditionnel dans les systèmes hypothétiques. », Annual Conference of the Linguistic Society of Belgium, Institut Libre Marie Haps (Brussels), 1997), et qui est le suivant.

Ce traitement consiste en l'occurrence à compter dans le document

- le nombre de propositions dans lesquelles le conditionnel présent sans condition est employé ;

- le nombre de propositions dans lesquelles le conditionnel hypothétique est employé.

Par exemple, pour une implémentation de l'outil sur des documents en anglais, l'outil déterminera :

- le nombre de propositions dans lesquelles le conditionnel présent sans condition est employé en comptant le nombre de propositions qui contiennent « would » et dont la suite ne contient pas d'auxiliaire infinitif ni de participe passé ;

- le nombre de propositions dans lesquelles le conditionnel hypothétique est employé en comptant le nombre de propositions subordonnées qui contiennent « if ».

Avantageusement, l'outil ajoute également à cette somme:

- le nombre de phrases,
- un paramètre fonction de la date de parution du document.

La somme obtenue est ensuite divisée par le nombre de phrases du document, de façon à disposer d'un résultat rapporté aux divisions naturelles du document que constituent les phrases.

La somme ainsi obtenue constitue un paramètre qui quantifie le taux de distanciation par document et qui traduit par conséquent un comportement linguistique du locuteur. Il permet une mesure graduelle entre les textes et permet de situer différents textes, les uns par rapport aux autres.

L'outil met ensuite en œuvre sur l'ensemble des documents traités une évaluation statistique : calcul du taux de distanciation moyen et d'un écart type de taux de distanciation.

Dans le cas où le traitement se fait sur une fenêtre temporelle glissante, on parle alors de taux de distanciation d'un flux de documents et de dispersion de taux de distanciation (écart type sur moyenne).

Les taux de distanciation moyens et les écarts types peuvent être comparés à des seuils et classés dans une catégorie de taux de distanciation selon le résultat de cette comparaison.

REVENDICATIONS

1. Outil pour le traitement d'au moins une partie d'un ensemble de documents textuels stockés dans une base de données, caractérisé en ce qu'il comporte des moyens pour déterminer automatiquement dans au moins un document le nombre de propositions dans lesquelles le temps conditionnel hypothétique et le nombre de propositions dans lesquelles le temps conditionnel présent sans condition et employé, et pour calculer un taux de distanciation qui est fonction du ou des nombres ainsi déterminés

2. Outil selon la revendication 1, caractérisé en ce que pour calculer le taux de distanciation du document, il comporte des moyens pour sommer le ou les nombres ainsi déterminés.

3. Outil selon la revendication 2, caractérisé en ce que pour calculer le taux de distanciation, il comporte des moyens pour sommer le ou les nombres ainsi déterminés et pour y ajouter le nombre de phrases et/ou un paramètre fonction de la date de parution du document.

4. Outil selon l'une des revendications précédentes, caractérisé en ce qu'il comporte des moyens pour déterminer le nombre de phrases du document et pour calculer un taux de distanciation rapporté à ce nombre de phrases.

5. Outil selon l'une des revendications précédentes, caractérisé en ce qu'il comporte des moyens pour calculer la moyenne et/ou l'écart type d'une pluralité de taux de distanciation calculés pour différents documents.

6. Outil selon la revendication 5, caractérisé en ce que sur la base de données évoluant dans le temps pour stocker de nouveaux documents, on sélectionne dans la base de données à plusieurs instants successifs les documents correspondant à une fenêtre temporelle que l'on fait glisser dans le temps et on détermine des moyennes et des écarts types de taux de distanciation pour ces différents instants.

7. Outil selon la revendication 6, caractérisé en ce qu'on compare une valeur moyenne et/ou un écart type de taux de distanciation à une valeur

seuil et on détecte l'instant où ladite valeur moyenne et/ou ledit écart type franchissent ledit seuil.